ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 24, April 2016

SPECTRAL CLUSTER BASED DECISION TREE DATA MINING TECHNIQUE FOR ANALYZING STUDENT PERFORMANCE IN HIGHER EDUCATION INSTITUTIONS

S.BHARATHI
PG Scholar
Computer Science And Engineering
Erode Sengunthar Engineering
College,Thudupathi

K.M.SUBRAMANIAM
Assistant Professor (SLG-II)
Computer Science And Engineering
Erode Sengunthar Engineering
College,Thudupathi

ABSTRACT

Data mining is the method of determining patterns in large datasets with artificial intelligence, machine learning, statistics and database systems. The main goal of higher education institution is to employ data mining methodologies for learning student's performance in the educations. Data mining offers many tasks that are used to analysis the student performance. The classification task is designed applied to calculate student's performance. In addition to, many approaches are used for data classification to support decision tree method. The decision tree technique is employed to accurately predict the student performance. The existing work presented a SVM Prediction technique for evaluating the student Grade Point Average (GPA) in computer education and instructional technology at the end of first, second, and third-year courses.

Three kinds of procedures are involved in SVM prediction data mining technique, that are data preparation, formulation of prediction model and evaluation of the SVM prediction model. By using linear arithmetic, the SVM prediction model performs tasks of classification and regression using linear combination of features based on variables. However, the SVM technique does not provide suggestive methods for enhancing the student GPA. This technique fails to determine the exact recollect values and does not carried out subjective analysis. To overcome these drawbacks, the proposed work presents Spectral Cluster based Decision Tree Data Mining Technique for analyzing Student Performance in Higher Education Institutions.

To develop Cluster based Decision Tree technique predicts exact student GPA with aiming at improves true positive values. In order to provide decision tree based suggestive methods which helps to increase the GPA of weak students. The performance measure of proposed technique are done with following metrics such as, Prediction Accuracy, True Positive, False Positive and Number of decision rules.

Key Terms: Data Mining, Student Performance, GPA, True Positive, False Positive.

1. INTRODUCTION

Data mining helps to extract the related information from the large and complex databases. Data mining techniques are applied to data investigation and data identifications. Two types of data analysis are classification and prediction that used to analysis extract predefined class models. Several classification techniques are available such as Decision tree algorithm, Bayesian network, neural network and Genetic algorithm that construct classification model. This classification model aids to predict the future leaning based on previous pattern. Data mining is also known as Knowledge Discovery in Database that represents extracting or “mining" knowledge from large volumes of data. Data Mining is used in academic field to improve student understanding of learning process for analyzing, extracting and estimating variables associated to the learning process of students.

Currently the large amount of information stored in academic database which include the positive information for predict of students performance. The main objective of higher education institutes is present quality education to improve student’s quality of executive decisions. The higher education system obtains maximum level of quality using discovering knowledge to calculate students’ performance. Information’s similar to number present, analysis, class and Assignment marks were collected from the student’s management system that used to determine the performance at the end of the semester.

2. LITERATURE SURVEY

In this paper **[1]** the author proposes a Latent Semantic Analysis to discover intrinsic structure of information by using organizing data objects into similarity groups or clusters. It is also called unsupervised learning due to class labels are does not indicating a priori partition of the objects. On the other head, supervised learning is a classification for which the data objects are previously identified with known classes.

This article **[2]** studies data structure investigation potentially using cluster analysis. Density structures are designed to apply class decomposition that improves performance of decision tree classifiers. By using cluster analysis, the Classes are decomposed and cluster merge estimation employs decision tree classifiers. After that measure of class decomposition is generated on C4.5 and CART classifiers.

In this paper **[3]**, the author identifies the issues large population speaker identification in noisy conditions. Major techniques for speaker identification are developed based on Mel-Frequency Cepstral Coefficients (MFCC), Gaussian Mixture Model (GMM) and Universal Background Model (UBM). The techniques are recognized to perform very well for small population identification in low-noise conditions.

The goal of this tutorial **[4]** is to present some perception on those questions. The different graph Laplacians are developed and presents their essential properties. In order to provide spectral clustering algorithms are derived from scratch via lots of different approaches.

In this paper **[5]**, the author develops a simple spectral clustering algorithm thought the Matlab. The algorithm is identified using devices on matrix perturbation theory and performs well. The author display efficient experimental results on a number of challenging clustering problems.

In this paper **[6]**, the author derives new cost functions for spectral clustering. It is derived from calculates of error between a given partition and a solution of the spectral to reduce normalized cut problem. The cost functions are reduced with respect to the partition results in spectral clustering algorithms. The learning algorithm is applied to the blind one-microphone speech separation problem, cost problem as one of segmentation of the spectrogram.

This paper **[7]** reviews a technique to synthesizing decision trees used in several systems and explain one system of ID3. A basic algorithm is described in this paper.

This article **[8]** published a Center for Research on the Education of

Students Placed at Risk (CRESPAR). A national research and development center is supported by a grant from the Institute of Education Sciences (IES). In addition, substance or opinions are expressed.

In this research **[9]**, the classification task is applied to estimate student's performance and several systems are used for data classification, the decision tree method. The knowledge is extracted to the students' performance at the end semester of examination. It useful in recognizing the dropouts and students require special attention with presents appropriate advising/counseling.

In this paper **[10]**, the author proposes a new cost function for spectral clustering based evaluate of error between a given partitions. The cost functions are decreased through partition, results in new spectral clustering algorithm.

The author develops tractable approximation is applied to cost function derived from the power method of computing eigenvectors.

In this paper **[11]** the author provides a basic account on multiclass spectral clustering. Initially, by using discrete clustering formulation, the continuous optimization problem is determined using Eigen decomposition. The author explains the role of eigenvectors as a generator of all optimal solutions through orthonormal transforms.

In this paper **[12]**, the author proposes two frameworks that integrate temporal smoothness in evolutionary spectral clustering. For both frameworks, the author creates with intuitions gained from the recognized k-means clustering problem. Then, designs equivalent cost functions used for the evolutionary spectral clustering problems.

In this paper **[13]**, the author extensively used spectral clustering algorithm in which clusters data by using eigenvectors of a similarity/affinity matrix. In order to calculate two critical problems in spectral clustering: (1) how to automatically determine the number of clusters, and (2) how to perform effective clustering given noisy and sparse data. An analysis of the characteristics of eigenspace is performed that generates (a) every eigenvectors of a data similarity matrix is not educational and relevant for clustering; (b) eigenvector selection is different due to uninformative/irrelevant eigenvectors used that produce poor clustering results; and (c) the equivalent eigenvalues are does not used for relevant eigenvector selection given a realistic data set.

In this paper **[14]** the author develops an efficient clustering algorithm for large scale graph data using spectral methods. The key concept is to generate a small number of "super nodes" connected to the regular nodes repeatedly. In order to, reduce the original graph into a sparse bipartite graph. By clustering the bipartite graph with spectral methods are increased efficiently except losing considerable clustering power.

In this paper **[15]** the author show how optimizing the Q function can be regenerated as a spectral relaxation problem and develops two new spectral clustering algorithms that request to maximize Q. Experimental results specify that the new algorithms are efficient and effective discovering good clustering's and the suitable number of clusters across a multiple real-world graph data sets.

3. SPECTRAL CLUSTER BASED DECISION TREE DATA MINING TECHNIQUE

This work proposes a Cluster based Decision Tree technique for Analyzing Student Performance in Higher Education Institutions. This technique used to predict exact student GPA with improved true positive values. Then, decision tree rule is applied to improve GPA of weak student. Spectral cluster based Decision Tree is introduced to organize the student GPA value at different intermediary exam result. Similar ranges of exam scores are performed into a set spectrum values for generating true positive and false positive values. Decision tree is created based on spectral cluster object similarity range threshold. Decision rules are invented to provide suggestive measure for improving GPA at graduation. Spectral exam score value objects with predominant factors are manipulate the rule formation.

This proposed technique is performed through higher education student data at 10 fold cross validation. The main advantages of proposed work to improve student GPA based on suggestive methods. True positive and false positive values are calculated to identify cluster object precision and recall. In addition to, Subjective analysis is carried out with predominant resultant GPA threshold.

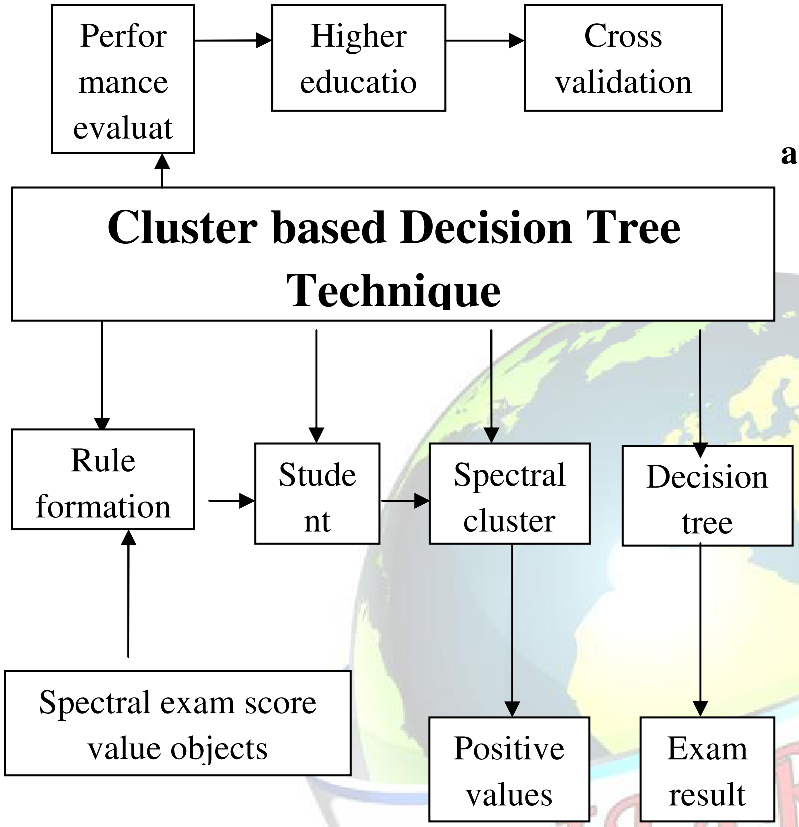

Fig: 3.1. Architecture Diagram of spectral cluster based decision tree data mining technique

The proposed technique is divided into three phases are:

a) Higher Educational Student Data Storage

b) Spectral Cluster for Student Data Organizing

c) Decision Tree on Student Activities and Prediction

a) Higher Educational Student Data Storage

The development of a higher educational institute is measured in terms of successful students of the institute. In higher education institute, student's academic performances are analyzed to predict the performances and it important requirement for improving quality education. Data mining techniques integrates most important role in data analysis with institutions collect the data to share from different data source stored at multiple student activities.

While student data storage is performed in different knowledge sources can be extracted from past and current data sets. Datasets contain scores of students which able to successfully completion of exam for their graduation. The student exam score are calculated by using standard evaluation system in a particular course. Student data stored in educational databases that comprises of multiple information in different perspective. In data mining technique, higher institutional student is used to evaluate factors influencing student performance and identify accurate view of information for specific student activity.

Multiple data are stored in institutional database that efficiently predict the course enrollment and estimate student failure rate in order to identify abnormal values in result sheets of students.

b) Spectral Cluster for Student Data Organizing

Spectral Cluster is classified the student data to evaluate the performance that efficiently analysis the changing parameters of student activity in the institution. Spectral cluster is much more useful in organizing GPA scores of student based on different (spectrum) changing metrics.

Initially, spectral cluster is constructed with standard threshold on same range of spectra. Then, Student exam marks with same range are organized into different spectra value. The instantaneous threshold is applied to evolve marking the spectra similarity range value for different parameters during changing student activity.

Based on the instantaneous threshold, the spectral clusters are regenerated with varied spectral cluster objects to calculate True positive and false positive values of the exam score.

c) Decision Tree on Student Activities and Prediction

Decision tree model is designed for identifying the source of current performance and predicting possible performance in forthcoming exams. It is formulated with spectral cluster object of instantaneous range threshold. Decision rules provide suggestion on parameters to be regulated for improving the student performance.

Spectral Cluster similarity rang objects refers to predominant factors for decision rule generation. For construction of decision tree model is used to identify multiple factors affecting student performance. Significant factors are considered from spectral cluster object for classifying students according to their attributes (grades).

Moreover, Cross validation and percentage split methods are developed to evaluate the efficiency of decision tree algorithm. From the decision tree and rules generated, the higher institution management provides appropriate actions to success rate of students increased efficiently.

4. PERFORMANCE EVALUATION

In this section evaluate the performance of Decision Tree Data Mining Technique for Analyzing Student Performance in Higher Education Institutions. One of the main contributions of the work is to predict Student Performance in Higher Education Institutions. Decision tree is constructed based on spectral cluster object similarity range threshold to provide suggestive measure for improving GPA at graduation. In order to calculate True positive and false positive values are identify cluster object precision and recall. The performance metrics of the parameters is Student Strength, prediction Accuracy, True Positive rate, False Positive rate, number of decision rules and Prediction Rate.

The performance metrics are

- No. of decision rules,
- Prediction Accuracy,
- True Positive rate,
- False Positive rate,
- Prediction Rate.

4.1 True Positive rate

The true positive rate is defined as the information's correctly identified as belonging to a specific class. It is measured in terms of %. The true positive rate is mathematically formulated as given below.

$$TPR = \left(\frac{\text{informations correctly identified as belonging to a class}}{c}\right)$$

Table: 4.1. No. of decision rules Vs True Positive rate (%)

No. of decision rules	**True Positive rate (%)**	
	SVMP(Existing)	**SC-DTDM(Proposed)**
10	13	15
20	16	18
30	19	23
40	22	27
50	25	31

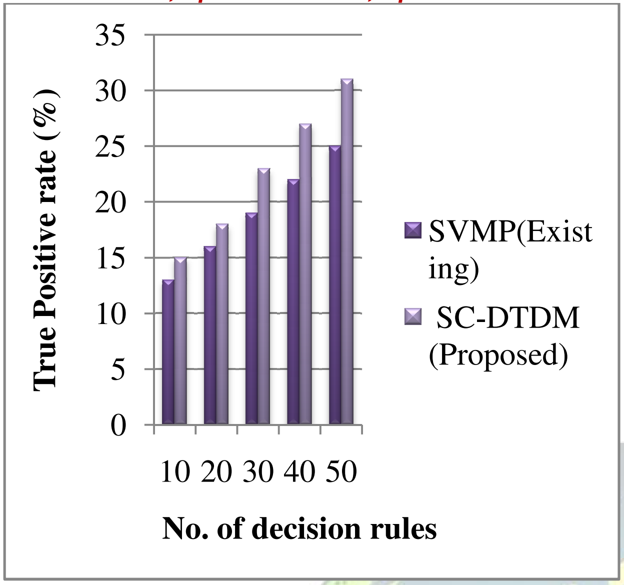

Figure: 4.1. No. of decision rules Vs True Positive rate (%)

Figure: 4.1. Demonstrate the rate of True Positive. X axis represents the No. of decision rules whereas Y axis denotes the True Positive rate using both the SVM based Prediction technique (SVMP) Technique and our proposed Spectral Cluster based Decision Tree Data Mining (SC-DTDM) Technique. When the No. of decision rules increased, True Positive rate get also increased accordingly. The rate of True Positive is illustrated using the existing SVMP and proposed SC-DTDM Technique. Figure 4.1.shows better performance of Proposed SC-DTDM method in terms of decision rules than existing SVMP and proposed SC-DTDM. The Spectral Cluster based Decision Tree Data Mining Technique achieves 15% high performance of True Positive rate variation when compared with existing system.

4.2 False positive rate

False positive rate is also called as false alarm rate the ratio of probability of wrongly analyzed as student performances to the total number of performance made. It is measured in terms of (%) and mathematically formulated given below,

$$FPR = \left(\frac{Probability\ of\ wrongly\ analyzed\ as\ student\ performaces}{No.of\ performances\ made}\right) * 100$$

Table: 4.2. No. of decision rules Vs False positive rate (%)

No. of decision rules	False positive rate (%)	
	SVMP(Existing)	SC-DTDM(Proposed)
10	62	56
20	66	59
30	70	63
40	73	67

50	77	69

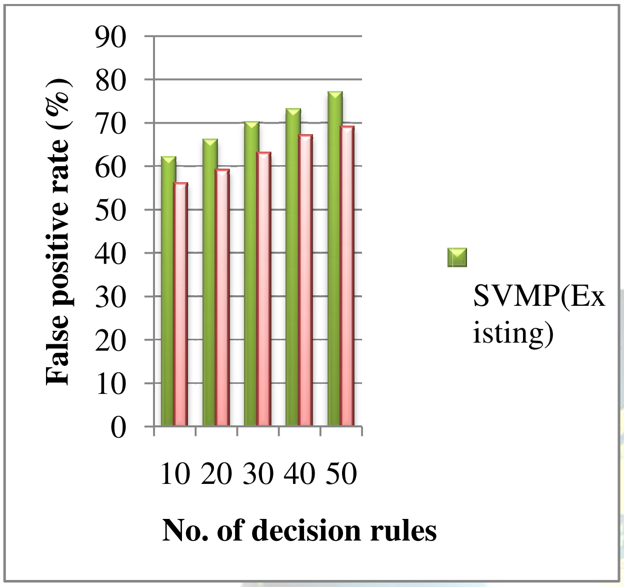

Figure: 4.2. No. of decision rules Vs False positive rate (%)

Figure: 4.2. Show the rate of False positive. X axis represents the No. of decision rules whereas Y axis denotes the false positive rate using both the SVM based Prediction technique (SVMP) Technique and our proposed Spectral Cluster based Decision Tree Data Mining (SC-DTDM) Technique. When the No. of decision rules increased, false positive get also reduced consequently. The rate of false positive is illustrated using the existing SVMP and proposed SC-DTDM Technique. Figure 4.2.shows better performance of Proposed SC-DTDM method in terms of decision rules than existing SVMP and proposed SC-DTDM. The Spectral Cluster based Decision Tree Data Mining Technique achieves 10% of False positive rate when compared with existing system.

4.3 Prediction accuracy

The higher educational domain of data mining is considered on evaluating the prediction accuracy of the educational performance of the students by using only influencing factors. The prediction accuracy of student score ranges using dataset that consists of all academic, personal and economic factors of the students.

Table: 4.3. No. of decision rules Vs prediction accuracy (%)

No. of decision rules	Prediction accuracy (%)	
	SVMP(Existing)	SC-DTDM(Proposed)
10	36	48
20	45	56
30	54	67
40	68	79
50	71	84

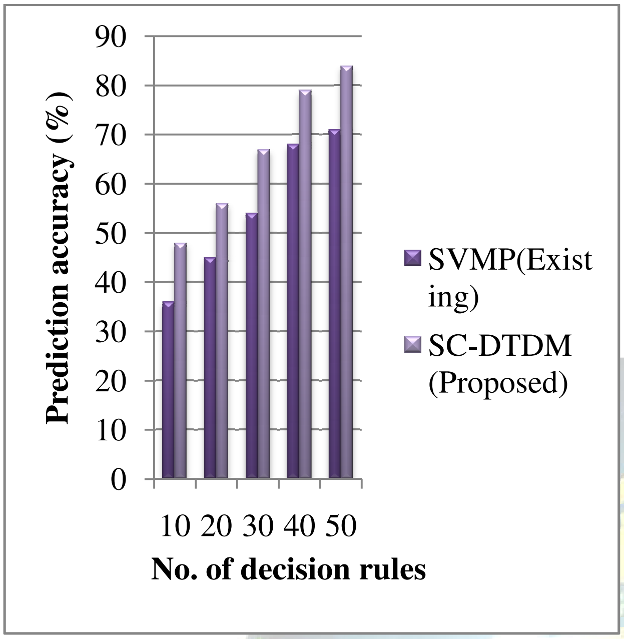

Figure: 4.3. No. of decision rules Vs prediction accuracy (%)

Figure: 4.3. Illustrate the prediction accuracy. X axis represents the No. of decision rules whereas Y axis indicates the prediction accuracy using both the SVM based Prediction technique (SVMP) Technique and our proposed Spectral Cluster based Decision Tree Data Mining (SC-DTDM) Technique. When the No. of decision rules increased, prediction accuracy gets also improved consequently. The prediction accuracy is demonstrated using the existing SVMP and proposed SC-DTDM Technique. Figure 4.3.shows better performance of Proposed SC-DTDM method in terms of decision rules than existing SVMP and proposed SC-DTDM. The Spectral Cluster based Decision Tree Data Mining Technique achieves 16% prediction accuracy when compared with existing system.

4.3 Prediction rate

The Proposed work performances are measured by prediction rate of student GPA. The prediction rate is used to perform exact predict the students similarity score ranges. It is measured in terms of (%).

Table: 4.4. No. of decision rules Vs Prediction rate (%)

No. of decision rules	Prediction rate (%)	
	SVMP(Existing)	SC-DTDM(Proposed)
10	40	45
20	45	49
30	49	53
40	53	58
50	59	65

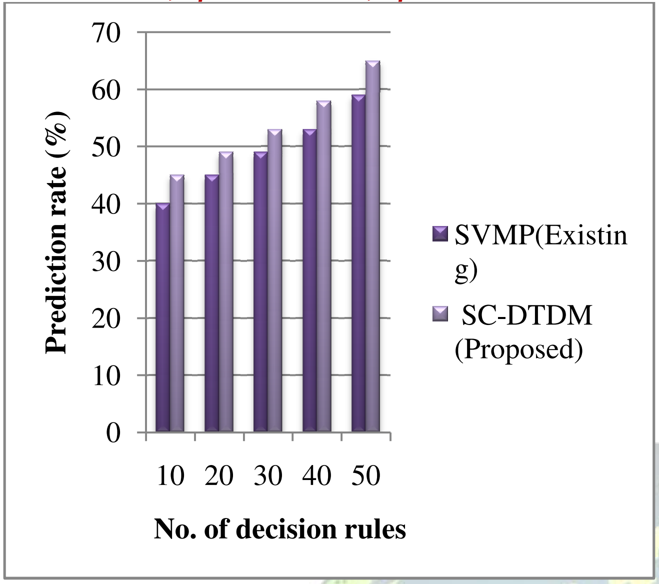

Figure: 4.4. No. of decision rules Vs Prediction rate (%)

Figure: 4.4. Show the rate of Prediction. X axis denotes the No. of decision rules whereas Y axis represents the Prediction rate using both the SVM based Prediction technique (SVMP) Technique and our proposed Spectral Cluster based Decision Tree Data Mining (SC-DTDM) Technique. When the No. of decision rules increased, Prediction rate get also increased consequently. The rate of Prediction is illustrated using the existing SVMP and proposed SC-DTDM Technique. Figure 4.4 displays better performance of Proposed SC-DTDM method in terms of decision rules than existing SVMP and proposed SC-DTDM. The Spectral Cluster based Decision Tree Data Mining Technique using Prediction rate improved by 8 % when compared with existing system.

5. CONCLUSION

This paper proposes Spectral Cluster based Decision Tree Data Mining Technique for Analyzing Student Performance in Higher Education Institutions. The Cluster based Decision Tree technique is applied to predict exactly student GPA with improved true positive values. The Spectral cluster is organized into student GPA value on different intermediary exam result. Subjective analysis is performed with predominant resultant GPA threshold. This in turn prediction accuracy is increased.

REFERENCES

[1] Bing Liu, Yiyuan Xia and Philip S. Yu, “Clustering Via Decision Tree Construction”, Proceedings of the ninth international conference on Information and knowledge management, Pages 1 – 25.

[2] Inese Polaka and Arkady Borisov, “Clustering-Based Decision Tree Classifier Construction”, Technological and economic development of economy, Volume 16, Issue 4, 2010, Pages 765 – 781.

[3] Yakun Hu, Dapeng Wu and Antonio Nucci, "Fuzzy-Clustering-Based Decision Tree Approach for Large Population Speaker Identification", IEEE Transactions on Audio, Speech, and Language Processing, Volume 21, Issue 4, Pages 1 – 13, April 2013.

[4] Ulrike von Luxburg, "A Tutorial on Spectral Clustering", Springer, Statistics and Computing, Volume 17, Issue 4, December 2007, Pages 1 – 32.

[5] Andrew Y. Ng, Michael I. Jordan and Yair Weiss, "On Spectral Clustering: Analysis and an algorithm", Advances in Neural Information Processing Systems, 8 Pages.

[6] Francis R. Bach and Michael I. Jordan, "Learning Spectral Clustering With Application to Speech Separation", Journal of Machine Learning Research, Volume 7, 2001, Pages 1963 – 2001.

[7] J. R. Quinlan, "Induction of decision trees", Machine Learning, Volume 1, Issue 1, 1986, Pages 81 – 106.

[8] Jeffrey C. Wayman, Sam Stringfield and Mary Yakimowski, "Software Enabling School Improvement through Analysis of Student Data", January 2004, 63 Pages

[9] Brijesh Kumar Baradwaj and Saurabh Pal, "Mining Educational Data to Analyze Students Performance", International Journal of Advanced Computer Science and Applications (IJACSA), Volume 2, Issue 6, 2011, Pages 63 – 69.

[10] Francis R. Bach and Michael I. Jordan, "Learning Spectral Clustering", June 2003, 8 Pages

[11] Stella X. Yu and Jianbo Shi, "Multiclass Spectral Clustering", Ninth IEEE International Conference on Computer Vision, 2003, Pages 1 – 7.

[12] Yun Chi, Xiaodan Song, Dengyong Zhou, Koji Hino and Belle L. Tseng, "Evolutionary Spectral Clustering by Incorporating Temporal Smoothness", Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, August 12 – 15, 2007, 10 Pages

[13] Tao Xiang and Shaogang Gong, "Spectral clustering with eigenvector selection", Elsevier, Pattern Recognition, Volume 41, Issue 3, March 2008, Pages 1012 – 1029.

[14] Jialu Liu, Chi Wang, Marina Danilevsky and Jiawei Han, "Large-Scale Spectral Clustering on Graphs", Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence, August 2013, Pages 1486 – 1492

[15] Scott White and Padhraic Smyth, "A Spectral Clustering Approach to finding Communities in Graphs", 2005, 12 Pages.